Common Forecast Verification Metrics Can Overestimate Skill

Thomas M. Hamill

NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

Josip Juras

Geophysical Institute, Faculty of Science, University of Zagreb

Zagreb, Croatia

Submitted to *Monthly Weather Review*

25 April 2005

Corresponding author address:

Dr. Thomas M. Hamill
NOAA-CIRES Climate Diagnostics Center
R/CDC 1, 325 Broadway
Boulder, CO 80301 USA

e-mail: tom.hamill@noaa.gov
phone: 1 (303) 497-3060

ABSTRACT

It is common practice to summarize the skill of weather forecasts using an agglomeration of samples spanning many locations and times. In many of these calculations, there is an implicit assumption that the climatological frequency of event occurrence is fixed for all samples. If the event frequency actually varies among the samples, then a fictitiously high skill may be reported. The extra fictitious skill reflects the ability of the forecast system to distinguish variations of the climatology among the samples rather than true forecast skill. This affects many common deterministic verification metrics such as threat scores. Probabilistic forecast metrics such as the Brier skill score, relative operating characteristic, and economic value diagrams are also affected. Demonstrations of the false skill are provided, and guidelines are suggested for how to adapt these diagnostics to avoid this problem.

1. **Introduction**

This article will demonstrate that many commonly used weather forecast verification metrics report positive forecast skill when none truly exists or report more skill than the forecast truly has. Depending on the metric and the event being verified, this effect can be large or small. Unfortunately, hundreds of peer-reviewed publications on weather forecast verification over the last half century may have reported exaggerated skill.

Our study of this problem has been motivated by our own experiences in weather forecast verification. We have encountered circumstances where we have diagnosed large positive skill when intuition suggested that little or no skill would exist. For example, the first author used a common probabilistic forecast verification metric, the relative operating characteristic, in a comparison of ensemble forecast methods (Hamill et al. 2000b). The author reported a relative operating characteristic curve for windspeed forecasts at 5 days lead that were consistent with a nearly perfect forecast, different than forecast experience would suggest. The second author discussed the overestimation of forecast skill (Juras 2000) in a comment on a Buizza et al. (1999) article. It was suggested that the chosen metrics might report false skill if climatological frequencies vary within the verification area. This issue has also been raised indirectly in other publications, including Buizza (2001; p. 2335), Atger (2003), and Glahn (2004; p. 770),

This article extends the comments of Juras (2000) and the others. We will examine four common skill metrics, the Brier skill score (Wilks 1995), the relative operating characteristic (Swets 1973, Harvey et al. 1992), economic value diagrams (Richardson 2000), and the equitable threat score (Schaefer 1990). All are capable of reporting positive forecast skill when none is present. Many other metrics such as the ranked probability skill score

(Wilks 1995, Epstein 1969, Murphy 1971) and other contingency-table based threat scores will not be discussed but are subject to the same problem.

Section 2 will provide a brief review of the four chosen verification metrics, as well as descriptions of how they are computed. Section 3 follows with a very simple example of false skill and an explanation of why it occurs. Section 4 shows that the false value may or may not be reported with real meteorological data, depending on what event is being considered. Section 5 demonstrates how large the effect can be for a common verification problem, the threat scores of short-range precipitation forecasts. Section 6 concludes with a discussion of the implications and how to change verification tactics to avoid this problem.

2. **Computation of common verification metrics**

Below, we review four general verification metrics, the Brier skill score, relative operating characteristic, economic value diagram, and equitable threat score. After the review, we describe how each of these metrics can be calculated in several different ways.

The long-used Brier score (Brier 1950) is a measure of the mean-square error of probability forecasts for a dichotomous (two-category) event, such as the occurrence/non-occurrence of precipitation. A review is provided in Wilks (1995), and references therein provide further background. The Brier score is often hard to interpret; is a Brier score of 0.06 good or bad? Consequently, the Brier score is often converted to a skill score, its value normalized by the Brier score of a reference forecast such as climatology (ibid). A Brier skill score (BSS) of 1.0 indicates a perfect probability forecast, while a BSS of 0.0 should indicate the skill of the reference forecast (see Mason 2004 for further discussion of whether a BSS of 0.0 indicates no skill).

The relative operating characteristic (ROC) has gained widespread acceptance in the past few years as a metric for ensemble verification. The ROC has been used for decades in engineering, biomedical, and psychology applications; see an overview in Swets (1973). Its application in meteorology was proposed in Mason (1982), Stanski et al. (1989), and Harvey et al. (1992). In the Hamill et al. (2000a) summary of an ensemble workshop, it was recommended by the ensemble verification community as a standard metric, and the ROC was recently made part of the World Meteorological Organization's (WMO) standard (WMO, 1992). Characteristics of the ROC have been discussed in Buizza et al. (1998), Mason and Graham (1999), Juras (2000), Wilson (2000), Buizza et al. (2000ab), Wilks (2001), Kheshgi and White (2001), Kharin and Zwiers (2003), and Marzban (2004). The technique has been used to diagnose forecast accuracy in, for example, Buizza and Palmer (1998), Buizza et al. (1999), Hamill et al. (2000b), Palmer et al. (2000), Richardson (2000, 2001ab), Wandishin et al. (2001), Ebert (2001), Mullen and Buizza (2001, 2002), Bright and Mullen (2002), Yang and Arritt (2002), Legg and Mylne (2004), Zhu et al. (2002), Toth et al. (2003), and Gallus and Segal (2004). Harvey et al. (1992) provide a thorough review of the concepts underlying the ROC.

Economic value diagrams were introduced to the meteorology community by Richardson (2000). These diagrams provide information about the potential economic value of ensemble forecasts for a particular event. The diagrams indicate the relative value as a function of the user's cost/loss ratio. A value of 1.0 indicates that the full economic value of a perfect forecast should be realized, and a value of 0.0 indicates the value of climatology. This framework was also used in Palmer et al. (2000) and

Richardson (2001b). Demonstrations of its application value can be found, for instance, in Richardson (2000), Palmer et al. (2000), Buizza et al. (2003), and Zhu et al. (2002).

The equitable threat score (ETS) provides one of many ways of summarizing the ability of a deterministic forecast to correctly forecast a dichotomous event. The ETS will produce a score of 1.0 for a perfect forecast, and random or climatological forecasts should be assigned a value of 0.0. The ETS is commonly used to evaluate the skill of forecasts, especially precipitation. See, for example, Rogers et al. (1995, 1996), Hamill (1999), Bayler et al. (2000), Stensrud et al. (2000), Xu et al. (2001), Ebert (2001), Gallus and Segal (2001), Chien et al. (2002), and Accadia et al. (2003).

The method for computing these metrics is now discussed, starting with the probabilistic metrics. The BSS, ROC, and economic value diagrams will be generated from ensemble forecasts, though they can be generated from any probabilistic forecast.

Start by defining a dichotomous event of interest, such as occurrence/non-occurrence of precipitation, or temperature above or below a threshold. Let $\mathbf{X}_{\mathrm{e}}(j,k) = [X_1(j,k), \ldots , X_n(j,k)]$ be an n-member ensemble forecast of the relevant scalar variable (again, precipitation or temperature) for the jth of m locations and the kth of r case days. The ensemble at that day and location is first sorted from lowest to highest. This sorted ensemble is then converted into an n-member binary forecast $\mathbf{I}_{\mathrm{e}}(j,k) = [I_1(j,k), \ldots , I_n (j,k)]$ indicating whether the event was forecast (=1) or not forecast (=0) in each member. The observed weather is also converted to binary, denoted by $I_o(j,k)$.

a. Brier skill scores

Assuming that each member forecast is equally likely, a forecast probability *p_f(j,k)* is calculated from the dichotomized ensemble:

$$p_f(j,k) = \frac{\sum_{i=1}^{n} I_i(j,k)}{n} \quad . \tag{1}$$

The Brier score of the forecast BS_f is calculated as

$$BS_f = \sum_{k=1}^{r} \sum_{j=1}^{m} \left(p_f(j,k) - I_o(j,k) \right)^2 \quad . \tag{2}$$

A Brier skill score (BSS) is calculated as

$$\text{BSS} = 1.0 - BS_f \,/\, BS_c \ , \tag{3}$$

where BS_c is the Brier score of the reference probability forecast, typically the probability of event occurrence from climatology.

An ambiguity and potential source of false skill may be traced to the method for calculating BS_c. One method would be to generate a climatological probability $p_c(j)$ of event occurrence unique to each location of the m locations in the domain,

$$p_c(j) = \frac{\sum_{k=1}^{r} I_o(j,k)}{r} \ , \tag{4}$$

in which case BS_c would be

$$BS_c = \sum_{k=1}^{r} \sum_{j=1}^{m} \left(p_c(j) - I_o(j,k) \right)^2 \quad . \tag{5}$$

Another way would be to calculate a climatology p_c averaged over all locations

$$p_c = \frac{\sum_{k=1}^{r} \sum_{j=1}^{m} I_o(j,k)}{r \cdot m} \ , \tag{6}$$

and let

$$BS_c = \sum_{k=1}^{r}\sum_{j=1}^{m}\left(p_c - I_o\left(j,k\right)\right)^2 \quad . \tag{7}$$

Differences in the calculation from using (4) – (5) instead of (6) – (7) will be illustrated in sections 3 and 4.

b. ROC diagrams

Calculation of the ROC starts with the population of 2x2 contingency tables, with separate contingency tables tallied for each sorted ensemble member and location. The contingency table for the *j*th location and *i*th sorted ensemble member has four elements: $\Gamma_i(j) = [\, a_i(j), b_i(j), c_i(j), d_i(j)]$, indicating the relative fraction of hits, misses, false alarms, and correct rejections (Table 1). The contingency table is populated using data over all *r* case days, and then each is normalized so the sum of the elements is 1.0.

The hit rate (*HR*) for the *i*th sorted forecast and *j*th location is defined as

$$HR_i\left(j\right) = \frac{a_i\left(j\right)}{a_i\left(j\right) + b_i\left(j\right)} . \tag{8}$$

Similarly, the false alarm rate is defined as

$$FAR_i\left(j\right) = \frac{c_i\left(j\right)}{c_i\left(j\right) + d_i\left(j\right)} . \tag{9}$$

The ROC for the *j*th of *m* locations is a plot of $HR_i\,(j)$ (ordinate) vs. $FAR_i\,(j)$ (abscissa), $i = 1, \ldots, n$. A ROC curve that lies along the diagonal *HR*=*FAR* line indicates no skill; a curve that sweeps out maximal area, as far toward the upper left corner as possible, indicates maximal skill.

It has often been judged to be more convenient to examine one rather than *m* different ROC curves. Hence, a single ROC is commonly generated from contingency tables averaged over all locations, i.e., $\Gamma_{\mathrm{i}} = \left(\bar{a}_{\mathrm{i}}, \bar{b}_{\mathrm{i}}, \bar{c}_{\mathrm{i}}, \bar{d}_{\mathrm{i}}\right)$ where $\bar{a}_{\mathrm{i}} = \sum_{j=1}^{m} a_{\mathrm{i}}(j) / m$, and $\bar{b}_{\mathrm{i}}, \bar{c}_{\mathrm{i}}$, and $\bar{d}_{\mathrm{i}}$ are similarly defined. Then

$$HR_i = \frac{\bar{a}_i}{\bar{a}_i + \bar{b}_i} \qquad (10)$$

and

$$FAR_i = \frac{\bar{c}_i}{\bar{c}_i + \bar{d}_i} \qquad (11)$$

c. Economic value diagrams

Table 1 also indicates the economic costs that are associated with each contingency. See Zhu et al. (2002) for a more complete review of the underlying principles. The assumption is that an economic decision may be made upon the forecast information. Suppose adverse weather is associated with the event $I_o(j,k)$=1. Based on the forecast information the decision maker can protect, at cost *C*, against adverse effects, taking an additional smaller unprotectable loss L_{u} if the event occurs. If the event is not forecast to occur but it does occur, a total loss $L = L_{\mathrm{p}} + L_{\mathrm{u}}$ is realized, where L_{p} is the additional loss that could have been protected against (L_{p} and L_{u} are commonly considered fixed quantities for a particular user). A correct NO forecast incurs no cost.

The expected expense due to a decision based on the *i*th ensemble member forecast at the *j*th location can be shown (ibid) to be

$$E_f(i,j,C) = a_i(j)\,(C + L_u) + c_i(j)\,C + b_i(j)\,(L_p + L_u) \quad . \tag{12}$$

Let $o(j)$ be the climatological frequency of the event occurrence, $o(j) = a_i(j) + b_i(j)$ (note that the same o(*j*) will be calculated regardless of the value of *i*). The expected expense associated with using climatological information for a decision is

$$E_c(C) = o(j)\,L_u + \text{Min}\left(o(j)\,L_p\,,\,C\right) . \tag{13}$$

The expected expense of a perfect forecast is

$$E_p\,(C) = o(j)\,(C + L_u) \,. \tag{14}$$

Assume L_p and L_u are fixed. The overall expected economic value for the *i*th sorted ensemble forecast at the *j*th location and the cost *C* is

$$V(i,j,C) = \frac{E_c(C) - E_f(i,j,C)}{E_c(C) - E_p(C)} \qquad . \tag{15}$$

This value is typically calculated for a range of *C* between 0 and L_p. At the *j*th location, the user then has *n* possible expected values associated with using each of the *n* ensemble forecasts as a possible decision threshold. The user typically chooses the one that provides the largest value.

$$V_{\max}(j,C) = \max\left(V(1,j,C)\,,\,\ldots\,,\,V(n,j,C)\right) \tag{16}$$

The determination of the optimal $V_{\max}(j,C)$ is typically re-calculated for other C's with values between 0 and L_p, since a different sorted ensemble member may provide the largest value for a different C. The optimal value is plotted as a function of C / L_p .

As with ROCs, the user may prefer to examine only one economic value diagram synthesizing information over all locations. This could be computed in two ways; an averaged value $\overline{V}_{\max}(C)$ could be computed first as an average of values at the different locations

$$\overline{V}_{\max}(C) = \frac{1}{m}\sum_{j=1}^{m} V_{\max}(j,C) . \tag{17}$$

Alternatively, economic value could be calculated from the average contingency tables Γ_i. In this case, $\overline{o} = \overline{a}_i + \overline{b}_i$, and then (12) – (14) are replaced by

$$\overline{E}_f\left(i,C\right) = \overline{a}_i\left(C + L_u\right) + \overline{c}_i C + \overline{b}_i\left(L_p + L_u\right), \tag{18}$$

$$\overline{E}_c(C) = \overline{o}L_u + Min(\overline{o}L_p, C) , \tag{19}$$

$$\overline{E}_p\left(C\right) = \overline{o}\left(C + L_u\right) . \tag{20}$$

Then (15) is replaced by

$$\overline{V}(i,C) = \frac{\overline{E}_c(C) - \overline{E}_f\left(i,C\right)}{\overline{E}_c(C) - \overline{E}_p(C)} . \tag{21}$$

(16) and (17) are replaced by

$$\overline{V}_{\max}\left(C\right) = \max\left(\overline{V}\left(1,C\right),\ldots,\overline{V}\left(n,C\right)\right) . \tag{22}$$

d. Equitable threat score

Assume now that we have a deterministic forecast rather than an ensemble. The ETS could be calculated for each *j* of the *m* locations using Table 1 (but dropping the $_i$ subscript denoting the ensemble member number). The equation for the *ETS* is

$$ETS(j) = \frac{a(j) - a_r(j)}{a(j) + b(j) + c(j) - a_r(j)}, \tag{23}$$

where $a_r(j)$ is the expected fraction of correct forecasts for a random forecast

$$a_r(j) = \frac{\big(a(j) + c(j)\big)\big(a(j) + b(j)\big)}{a(j) + b(j) + c(j) + d(j)}. \tag{24}$$

Commonly the ETS is calculated using contingency tables summed over all the grid points. Let $\bar{a} = \sum_{j=1}^{m} a(j)/m$, and define $\bar{b}, \bar{c}$, and $\bar{d}$ similarly. Then an ETS that presumably represents the domain-averaged skill is calculated from

$$ETS = \frac{\bar{a} - \bar{a}_r}{\bar{a} + \bar{b} + \bar{c} - \bar{a}_r}, \tag{25}$$

where

$$\bar{a}_r = \frac{\left(\bar{a} + \bar{b}\right)\left(\bar{a} + \bar{c}\right)}{\bar{a} + \bar{b} + \bar{c} + \bar{d}}. \tag{26}$$

3. **An example of false skill: synthetic data at two independent locations**

Suppose our world consists of two small, isolated islands, and suppose weather forecasting is utterly impossible on this planet; the best one can do is to forecast the climatological probability distribution appropriate to each island. To simulate this, assume that at island 1, the daily maximum temperature was randomly sampled from its climatological distribution ~ N(+2, 1), that is, the temperature was a draw from a normal distribution with a mean of 2.0 and a standard deviation of 1.0. At island 2, the daily maximum temperature ~ N(-2, 1). 100-member ensembles of weather forecasts were generated by taking random draws from each island's climatology. 100,000 days of weather and ensemble forecasts were simulated, and we consider the event that the temperature was greater than 0. On island 1, both verification and ensemble ~ N(+2, 1) and were drawn independently. The same process was repeated for island 2, but verification and ensemble ~ N(-2, 1) .

a. Brier skill scores

From the synthetic verification and sorted ensembles, the BSS was calculated two ways, assuming the reference score could be calculated individually using (5), or over both islands using (7). The BSS was 0.0 (correct) when using (5) and 0.95 (incorrect) when using (7). Using a climatology averaged over the two stations as the reference was clearly inappropriate.

b. Relative operating characteristics

ROCs were generated for each island individually (Figs. 1 a–b) using (8) - (9), and indeed, these each show no skill (area = 0.5). To generate one ROC over the two islands, (10) – (11) were used. A ROC was then generated from the pooled tables (Fig. 1c). Note the very large positive area under the ROC curve, suggesting nearly perfect forecast skill.

Why was skill now indicated by the ROC? By compositing data over the two islands, the ROC analysis no longer implicitly assumed that the climatological distribution was ~ N(+2, 1) ***or*** ~ N(-2, 1). *Rather, it assumed that the climatological distribution was* ~ 0.5 • N(+2, 1) +0.5 • N(-2, 1), *a bimodal distribution. Further, the contingency tables were populated consistent with the assumption that the forecast perfectly predicted which mode of the distribution the verification lay in*; when the forecasts were drawn from the positive mode N(+2, 1), the observed states were also drawn from the positive mode N(+2, 1), and when the forecasts were drawn from N(-2, 1), the observed state were drawn from N(-2, 1) as well. This can be demonstrated by generating a ROC simulated under these assumptions. Such a ROC is identical to that in Fig. 1c. This illustrates that the ROC can report false skill in situations where the climatologies differ among the samples used to populate the contingency tables. The ROC credits a forecast with having skill merely if the sample's climatology are different than the climatology of the sum of the samples.

c. Economic value diagrams.

Figure 2 shows the economic value diagrams under the assumption that $L_u = 0$. As with the ROCs, the economic value was nil when computed at the individual islands using (16), but the diagram indicated that when averaged contingency tables and (22) were used, near-perfect economic value was realized at moderate cost/loss ratios. The underlying explanation is the same as for the ROC, the redefinition of climatology from the inappropriate compositing of contingency table elements.

d. Equitable threat score

The ETS for islands 1 and 2, calculated using (23), are 0.0 at each island. When contingency tables are added and (25) is used to calculate ETS, the score is 0.86.

e. Synthesis

Suppose we had sampled not from N(+2, 1) and N(-2, 1) distributions but from N(α, 1) and N(-α, 1), where α could be arbitrarily changed. If α were 0, then of course, no false skill would be reported, for the two islands would have the same climatology. As α is increased, we would expect to see an increase in the amount of false skill. Figure 3 illustrates this, repeating the experiment above and plotting the area under the ROC curve using eqs. (10) – (11), the BSS using (7), and the ETS using (25), as a function of α. The more the climatologies differ between the samples, the larger the false skill reported.

4. **Climatological forecasts of 850 hPa temperature**

Consider whether or not false skill can be reported with real data. 0000 UTC 850 hPa temperature analyses were extracted from the NCEP-NCAR reanalysis at a set of 26x12 grid points covering the conterminous United States (US). Data was considered for the first ~ 2 months (60 days) of 1979 to 2001. The grid spacing was 2.5° in latitude and longitude. Let T denote the temperature at a grid point, and T' denote the temperature anomaly from the mean. Three events were considered: (1) $T > 0$C, (2) $T' > 3$C, and (3) $T' > Q_{2/3}$, where $Q_{2/3}$ was the upper tercile of the climatological distribution, i.e., the temperature threshold defining the boundary between the lower twothirds of the distribution and the upper third. $Q_{2/3}$ was specified uniquely for each grid point.

First the method for generating contingency tables for the event $T > 0$C is described. For each of the first 60 days of the year and for each of the 23 years (1380 samples), the following process was performed at each grid point: (1) the analyzed temperature was extracted at that grid point, (2) the cross-validated, climatological probability of the event was determined using the other 22 years of data, (3) a cross-validated, 50-member ensemble was randomly drawn from the other 22 years of temperature samples at that grid point, (4) the ensemble was sorted, and (5) contingency tables were populated for that grid point. After all grid points were processed in this manner, average contingency tables for all of the grid points were also generated. To generate contingency tables for the ETS, the process was the same, but a single random sample from the climatology was drawn rather than an ensemble.

When generating ROCs, economic values, and ETSs for the events $T' > 3$C, and $T' > Q_{2/3}$, several additional steps were required. After step (1) above, the climatological mean for each date and location was determined and subtracted from the temperature, creating a database of temperature anomalies. The estimated climatological mean was estimated using a 30-day window centered on each day and cross-validated by year, using the remaining 22 years. Also, the terciles of the distribution were determined for each grid point.

a. T > 0 C

The climatological probabilities for this event varied from 0.005 in the north to

1.0 in the south. The mean climatological probablility was 0.59 with a standard deviation of 0.36.

When a location-dependent reference climatology was used (eqs. 4-5), the BSS was -0.03. When a domain-averaged climatology was used (eqs. 6-7), the BSS reported a false skill of +0.52.

Figure 4a shows ROCs calculated from the individual grid point data; the ROC for every third grid point in the N-S and E-W directions are plotted. The ROCs exhibit sampling variability but lie close to the HR=FAR line. However, the ROC based on a contingency table summed up over all the grid points (Fig. 4b) diagnosed a very large amount of skill. Figure 4c shows that when the economic value is calculated separately at each grid point and then averaged, its value was effectively zero. However, the economic value calculated from the contingency table sums was large. Again, these were artifacts of the widely differing climatologies for the grid points, as in section 2.

Table 2 reports the ETS for this event. The ETS was calculated for each of the *m* locations using (23) and then averaged. For some of these locations, the denominator of (23) was zero and the ETS was undefined, so the average ETS reported in Table 1 was calculated excluding these locations. The ETS was also calculated using the summed contingency tables and (25), excluding the same locations in calculating the table sums. As Table 2 shows, the average ETS was zero, but the ETS from the table sums was 0.345, reporting a false positive skill because of the differing climatology.

b. T ' > 3 C

Considering events defined by anomalies of temperature rather than temperature itself, the ensemble should have a much more consistent climatology from grid point to grid point. However, at the southernmost, more tropically influenced grid points, a deviation of 3C represented a relatively large deviation from climatology, while at the northernmost grid points, 3C reported a smaller deviation. The climatological probability of exceeding a 3C deviation ranged from 0.44 in the north to 0.07 in the south. The mean climatological probability was 0.30 with a standard deviation of 0.08.

When the location-dependent reference climatology was used, the reported BSS was -0.03. When the domain-averaged climatology was used, the BSS was -0.002. The extra skill when using the domain-averaged climatology was much less than when the fixed threshold was tested in section 4a; this was a consequence of the climatological probabilities varying much less widely.

Figures 5 a-b show the ROCs for individual grid points and from the summed contingency tables, respectively, and Fig. 5c shows the economic values as in Fig. 4c. The area under the ROC curve was much reduced but was still slightly greater than the expected 0.5. The economic value from the contingency table sums still reported unrealistic positive value at cost-loss ratios around 0.3, but they were much smaller.

The ETS reported in Table 2 increased only a bit more than 1 percent when changing from reporting the average of the grid points to contingency table sums.

c. $T' > Q_{2/3}$

By evaluating the probability of exceeding a quantile of the distribution, the climatological probabilities have been rendered uniform across all grid points; the climatology probability is of course 1/3 for this event. By construction, the BSS was the same for both, -0.03 (it was less than zero because the 50-member random draw from climatology only approximates the true climatology). With ROCs and economic values, whether we examined the average of scores at the grid points or computed the scores from contingency table sums, we found no skill (Fig. 6). Similarly, the ETSs (Table 2) reported the same lack of skill regardless of the how the ETS was computed.

5. **Equitable threat scores for numerical precipitation forecasts**

One of the primary goals of the U. S. National Weather Service is to improve forecasts of precipitation. The ETS is one measure that is very commonly used to evaluate the skill of their deterministic forecasts. The most common approach is to estimate the ETS for fixed precipitation thresholds from a contingency table populated over many days or months and over a wide geographic region such as the conterminous US. We demonstrate here that the ETS calculated in this manner can drastically overestimate forecast skill.

To demonstrate this, a very large set of numerical forecasts was used, provided by the analog forecast technique discussed in Hamill et al. (2005). The details of the forecast methodology can be found in this reference but are not particularly important here. What is germane is that we produced a 25-year time series of gridded deterministic precipitation forecasts, all using the same model and forecast technique. These forecasts have characteristics similar to those of current operational forecasts. For this demonstration, we limit ourselves to considering the ETS of the mean of the ensemble of analog forecasts over the conterminous US for January and February from 1979 to 2003. Both the forecast and the verification data (from the North American Regional Reanalysis, Mesinger et al. 2005) are on a ~30 km grid.

Figure 7 illustrates the geographic dependence of the ETS on forecast location. Skill is much larger in the southeast US and along the west coast than in the northern Great Plains. Table 3 provides the ETS, calculated both as an average of the values at the grid points (eq. 23) and from the contingency table sums (eq. 25). Notice that skill from contingency table sums is increasingly overestimated as the precipitation threshold is increased. The same underlying problem is at work; it is possible to take two different contingency tables that each report an ETS of zero, sum them, and report a positive ETS. The greater the relative differences in the climatology, the greater the relative effect.

6. **Discussion**

The preceding examples have demonstrated that the Brier skill score, relative operating characteristic, economic value diagrams, and the equitable threat score must be used with care when verifying weather forecasts. Typically, the meteorological question being asked is something akin to “what is the general skill of my forecast averaged over Europe?” The naïve approach for calculating the Brier skill score may be to compute it under the assumption that the climatology is invariant across the verification region. Similarly, when diagnosing the relative operating characteristic, economic value, or equitable threat score, a common step is to composite the forecast data into contingency tables that accumulate weather information across the domain. The preceding analysis showed that these diagnostics may falsely report positive skill in situations where the climatology differs across the domain. The more the climatology differs, the larger the falsely reported skill. By logical extension, false skill may also be reported if the verification samples span different seasons or even different times of the day with different climatologies.

Several implications can be made about forecast verification:

• Many prior verification studies (including at least two by the lead author, Hamill 1999 and Hamill et al. 2000b), should be re-evaluated, for the reported skill may be erroneous.

• In order to avoid reporting false skill, the researcher can alter his or her verification methodology. Alternative methodologies can be used that should not report false skill, such as: (1) analyze events where the climatological probabilities are the same throughout the sample (e.g., Buizza et al. 2003, Fig. 5, or Zhu et al. 2002). Section 4 demonstrated that, for example, relative operating characteristics, economic value diagrams, and equitable threat scores of climatological forecasts of *quantiles* of the 850 hPa temperature distribution did not report false positive skill. Regardless of whether the climatological means and variances are large or small, the fraction events classified as "yes" events are identical for different locations or times of the year. (2) If sample sizes are large enough, perform the calculations separately each for sub-sample with a different climatology. The data could then be summarized in some manner; for the relative operating characteristic, perhaps with a histogram of area for each of the sub-samples; for the equitable threat score or Brier skill score, perhaps with a histogram or a map of the skill scores at individual grid points with differing climatologies (e.g., Fig. 7); for economic value, perhaps with an average of the value curves at individual grid points.

- The *specific details* regarding how the verification metrics are calculated should be fully described in journal articles and texts. Minor changes in the methodology can dramatically change the reported scores.

- Other scores such as the ranked probability skill score (Wilks 1995) can also falsely report positive skill, just as with the Brier skill score. Whatever the chosen verification metric, it is wise to verify that climatological forecasts give the expected no-skill result before proceeding.

- Richardson (2001) demonstrated in a carefully controlled experiment that there was a theoretical equivalence between the Brier skill score and the integral of economic value assuming that users have a uniform distribution of cost-loss ratios between 0 and 1. One of the underlying assumptions was an invariant climatology across all samples. If this assumption is not met, then neither is this equivalence.

Despite these specific recommendations, we hope readers understand a more worrying implication: *we have forgotten or ignored the assumptions underlying the correct application of many weather verification techniques*. For example, the relative operating characteristic can be traced back to its roots in biostatistics and engineering literature. The underlying theory assumes that the tester seeks information on the differences between two fixed distributions. Perhaps the tester seeks to describe differences in the distribution of blood pressure for a group on a particular medication and a distribution for a control group not on the medication. With two fixed distributions, the relative operating characteristic is able to quantify the tradeoffs between Type I statistical errors (inappropriate acceptance of the null hypothesis) versus Type II statistical errors (inappropriate rejection of the alternative hypothesis) as a decision threshold is changed. What is the meaning of the relative operating characteristic when contingency tables are comprised of samples from distributions that are not fixed? Certainly, we cannot expect it to tell us about tradeoffs between Type I and Type II errors, for we have violated the underlying assumptions.

We hope this article will stimulate others to re-examine forecast verification. Do we commonly violate other underlying assumptions?

Acknowledgments

Dan Wilks and Matt Briggs (Cornell University), Craig Bishop (Navy/NRL), Beth Ebert (Australian BOM), Steve Mullen (U. Arizona), Simon Mason (IRI), Bob Glahn (NOAA/MDL), Neill Bowler and Ken Mylne (UK Met Office), Bill Gallus (Iowa State), Frederic Atger (Meteo France), Francois LaLaurette (EMCWF), Zoltan Toth (NCEP), and Jeff Whitaker (NOAA/CDC) are thanked for their discussions and comments on an early version of this manuscript. This research was supported by National Science Foundation grants ATM-0130154 and ATM-0205612.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918-932.

Atger, F., 2003: Spatial and interannual variability of reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509-1523.

Bayler, G. M., R. M. Aune, and W. H. Raymond, 2000: NWP cloud initialization using GOES sounder data and improved modeling of nonprecipitating clouds. *Mon. Wea. Rev.*, **128**, 3911–3920.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1-3.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the southwest monsoon. *Wea. Forecasting*, **17**, 1080–1100.

Buizza, R., and T. N. Palmer. 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518.

----------, T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **124**, 1935-1960.

----------, A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168-189.

----------, --------- , ---------, and --------, 2000a: Reply to comments by Wilson and by Juras. *Wea. Forecasting,* **15**, 367-369.

----------, J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000b: Current status and future development of the ECMWF ensemble prediction system. *Meteor. Appl*., **7**, 163-175.

----------, 2001: Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev*., **129**, 2329-2345.

----------, D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble prediction system and comparison with poor-man's ensembles. *Quart. J. Royal Meteor. Soc*., **129**, 1269-1288.

Chien, F.-C., Y.-H. Kuo, and M,-J. Yang, 2002: Precipitation forecast of MM5 in the Taiwan area during the 1998 Mei-yu season. *Wea. Forecasting*, **17**, 739–754.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev*., **129**, 2461–2480.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor*., **8**, 190-198.

Gallus, W. A., Jr., and M. Segal, 2001: Impact of improved initialization of mesoscale features on convective system rainfall in 10-km Eta simulations. *Wea. Forecasting*, **16**, 680–696.

------------, and -----------, 2004: Does increased predicted warm-season rainfall indicate enhanced likelihood of rain occurrence? *Wea. Forecasting*, **19**, 1127–1135.

Glahn, B., 2004: Discussion of verification concepts in "Forecast Verification: A Practitioner's Guide in Atmospheric Science." *Wea. Forecasting*, **19**, 769-775.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

------------ , C. Snyder, D. P. Baumhefner, Z. Toth, and S. L. Mullen, 2000a: Ensemble forecasting in the short to medium range: report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653-2664.

------------ , ------------- , and R. E. Morss, 2000b: A comparison of probabilistic forecast from bred, singular vector and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835-1851.

------------ , and J. S. Whitaker, 2005: Reforecasts, an important new data set for improving weather predictions. Bull. Amer. Meteor. Soc., submitted. Available at http://www.cdc.noaa.gov/people/tom.hamill/reforecast_bams.pdf .

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863-883.

Juras, J., 2000: Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Wea. Forecasting,* **15**, 365-366.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150.

Kheshgi, H. S., and B. S. White, 2001: Testing distributed parameter hypotheses for the detection of climate change. *J. Climate*, **14**, 3464–3481.

Legg, T. P., K. R. Mylne. 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906.

Marzban, C. 2004: The ROC curve and its area under it as performance measures. *Wea. Forecasting*, **19**, 1106-1114.

Mason, I. 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291-303.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713-725.

-----------, and -----------, 2002: Areas beneath relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Quart. J. Royal Meteor. Soc.*, **128**, 2145-2166.

----------- , 2004: On using “climatology” as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891-1895.

Mesinger, F., and coauthors, 2005: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, submitted.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638–663.

-----------, and ----------, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155-156.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Royal Meteor. Soc.*, **126**, 2013-2033.

Richardson, D. S. , 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **126**, 649-667.

-----------, 2001a: Ensembles using multiple models and analyses. *Quart. J. Royal Meteor. Soc.*, **127**, 1847-1864.

-----------, 2001b: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Royal Meteor. Soc.*, **127**, 2473-2489.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational "early" Eta Model: original 80-km configuration and recent changes. *Wea. Forecasting*, **10**, 810-825.

------------, and coauthors, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting*, **11**, 391-413.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989. *Survey of common verification methods in meteorology*. Enviroment Canada Research Report 89-5, 114 pp.

Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev*., **128**, 2077-2107.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science*, **182**, 990-999.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Chapter 7 of "*Forecast Verification: A Practitioner's Guide in Atmospheric Science*." John Wiley and Sons, 254 pp.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks. 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev*., **129**, 729–747.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Cambridge Press. 547 pp.

-----------, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8**, 209-219.

Wilson, L. J., 2000: Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Wea. Forecasting,* **15**, 361-364.

World Meteorological Organization, 1992: *Manual on the Global Data Processing System*, section III, Attachment II.7 and II.8, (revised in 2002). Available from http://www.wmo.int/web/www/DPS/Manual/WMO485.pdf.

Xu, M., D. J. Stensrud, J.-W. Bao, and T. T. Warner, 2001: Applications of the adjoint technique to short-range forecasting of mesoscale convective systems. *Mon. Wea. Rev.*, **129**, 1395-1418.

Yang, Z., and R.W. Arritt, 2002: Tests of a perturbed physics ensemble approach for regional climate modeling. *J. Climate*, **15**, 2881–2896.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. R. Mylne. 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

LIST OF TABLES

Table 1: Contingency table for the *i*th of the *n* sorted members at the *j*th location, indicating the relative fraction of hits [a_i(j)], misses [b_i(j)], false alarms [c_i(j)], and correct rejections [d_i(j)]. The economic costs associated with each contingency are also shown and are discussed in the text.

Table 2: Equitable threat scores for the events $T > 0$, $T' > 3$, and $T' > Q_{2/3}$, calculated as an average over all grid points with nonsingular ETS, and the ETS from the sum of contingency table elements at these grid points

Table 3: Equitable threat scores for the events of 1-2 day precipitation forecast amount greater than 1, 5, 10, 25, and 50 mm. ETS calculated as an average over all grid points with nonsingular ETS, and the ETS from the sum of contingency table elements at these grid points.

LIST OF FIGURES

Figure 7: ETS for 1-2 day 1 mm precipitation forecasts as a function of location, using Jan-Feb 1979-2003 forecast and observational data.

		Event forecast by *i*th member?	
		YES	NO
Event Observed?	YES	a_{i}(j) Mitigated loss (C+L_{u})	b_{i}(j) Loss ($L = L_{\mathrm{p}} + L_{\mathrm{u}}$)
	NO	c_{i}(j) Cost (C)	d_{i}(j) No cost

Table 1: Contingency table for the *i*th of the *n* sorted members at the *j*th location, indicating the relative fraction of hits [a_{i}(j)], misses [b_{i}(j)], false alarms [c_{i}(j)], and correct rejections [d_{i}(j)]. The economic costs associated with each contingency are also shown and are discussed in the text.

	Event		
	$T > 0$	$T' > 3$	$T' > Q_{2/3}$
ETS (average of grid points)	-0.001	-0.001	-0.002
ETS (contingency table sum)	0.345	0.012	-0.002

Table 2: Equitable threat scores for the events $T > 0$, $T' > 3$, and $T' > Q_{2/3}$, calculated as an average over all grid points with nonsingular ETS, and the ETS from the sum of contingency table elements at these grid points.

	Event				
	> 1mm	*> 5mm*	*> 10mm*	*> 25mm*	*> 50mm*
ETS (avg. of grid points)	0.362	0.264	0.162	0.019	0.002
ETS (contin. table sum)	0.426	0.438	0.369	0.115	0.041

Table 3: Equitable threat scores for the events of 1-2 day precipitation forecast amount greater than 1, 5, 10, 25, and 50 mm. ETS calculated as an average over all grid points with nonsingular ETS, and the ETS from the sum of contingency table elements at these grid points.

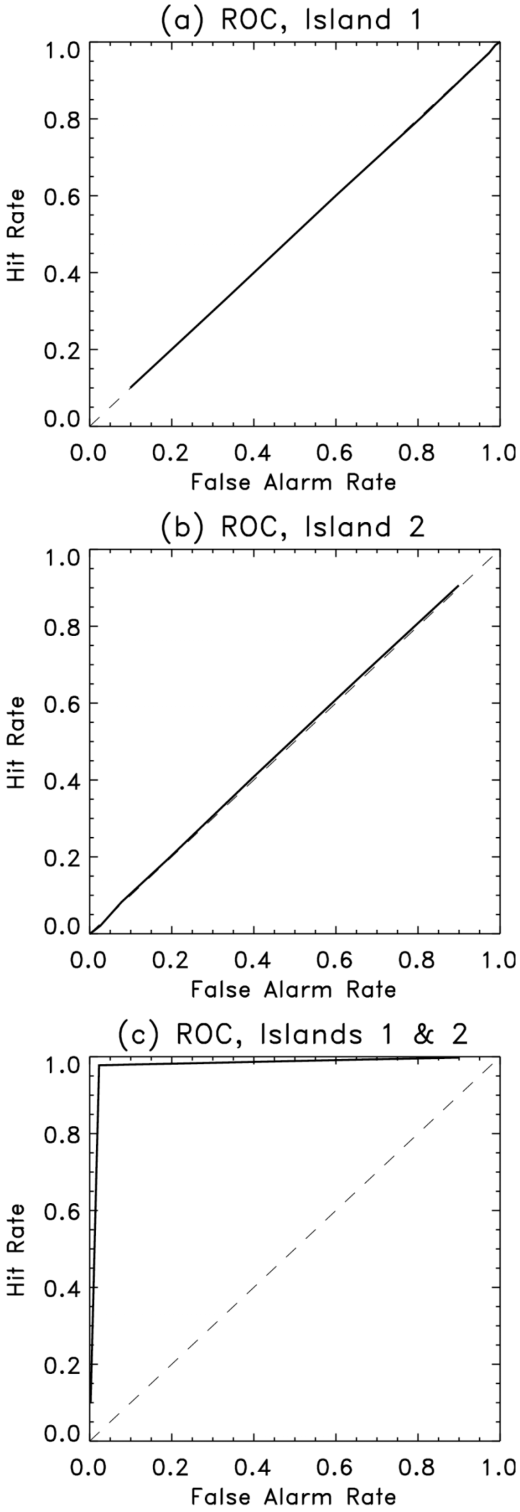

Figure 1: ROC diagrams for the event of temperature > 0. (a) Island 1, (b) Island 2, (c) Islands 1 and 2 together.

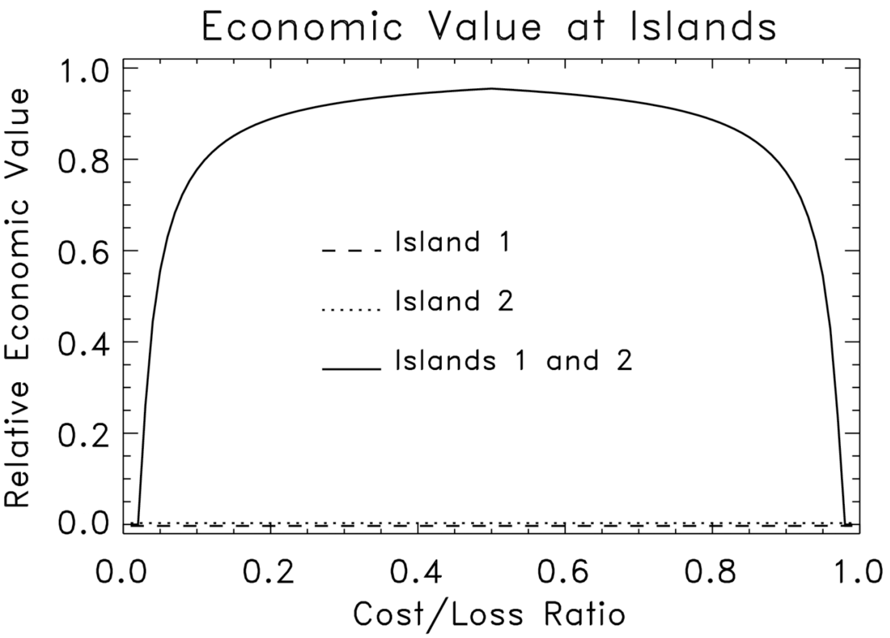

Figure 2: Economic value for the event temperature > 0 at islands 1, 2, and both.

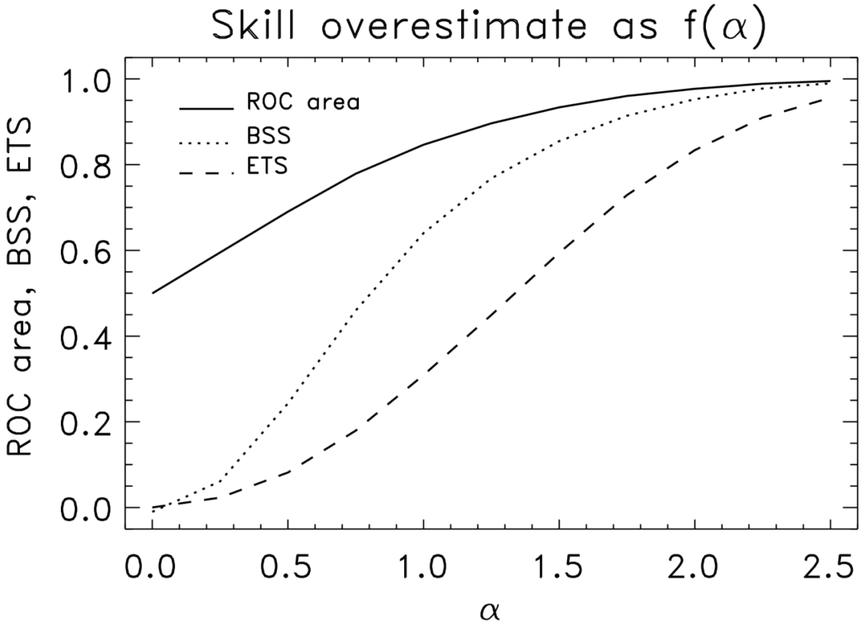

Figure 3: ROC area, BSS, and ETS as a function of the parameter α describing the difference in the means of the distributions between the two islands. Skill scores are calculated assuming a composite climatology.

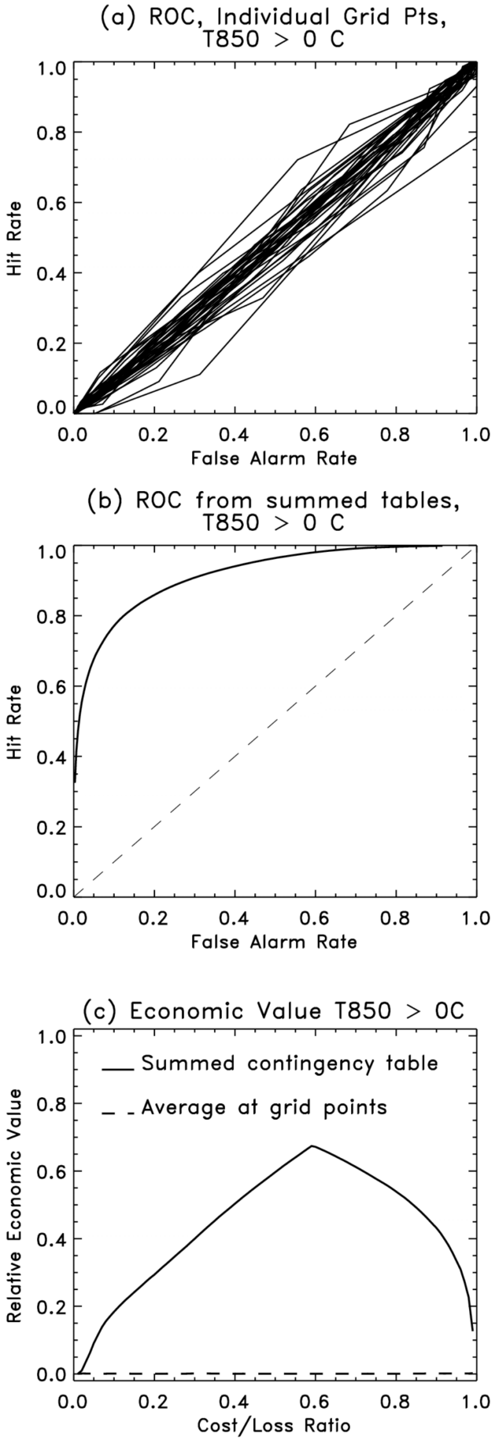

Figure 4: ROC and economic value for the event of 850 hPa temperature > 0 C using random draws from climatology using data from January-February 1979-2001. (a) ROC curves for selected individual locations around conterminous US, (b) ROC curve based on sum of contingency tables at individual grid points, and (c) economic value, plotted both as an average of values at individual grid points (dashed), or from the contingency table sums (solid).

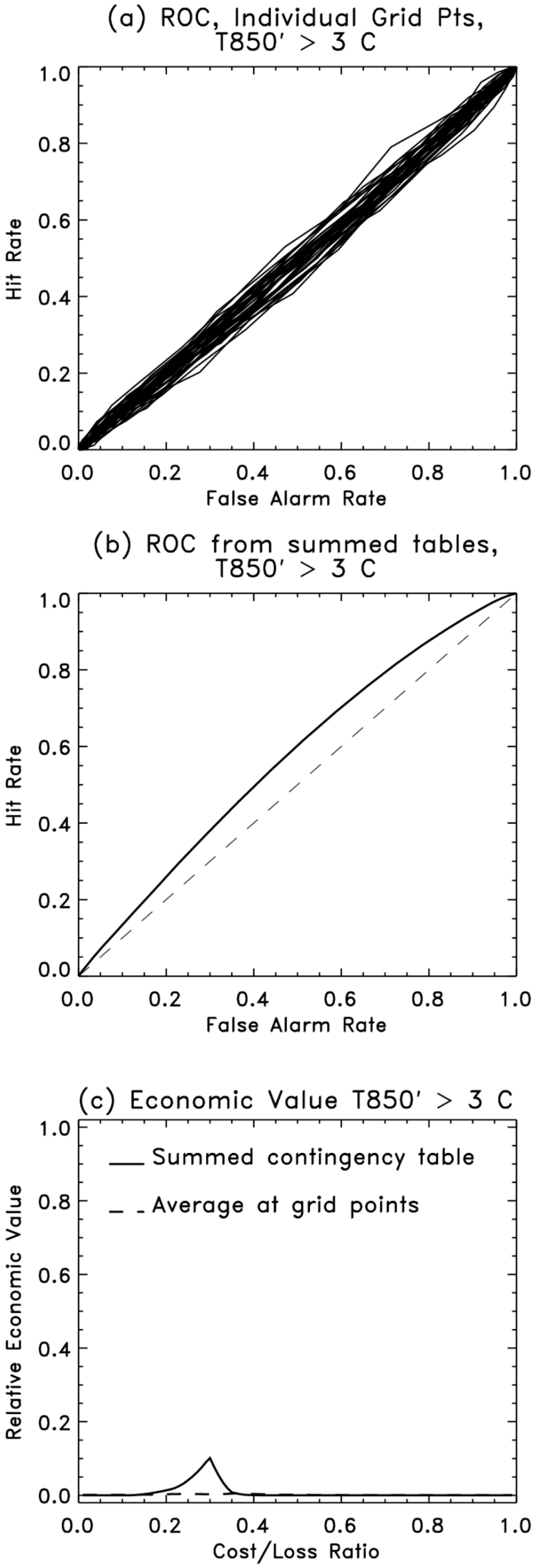

Figure 5: As in Fig. 4, but for the event of 850 hPa temperature anomaly > 3C.

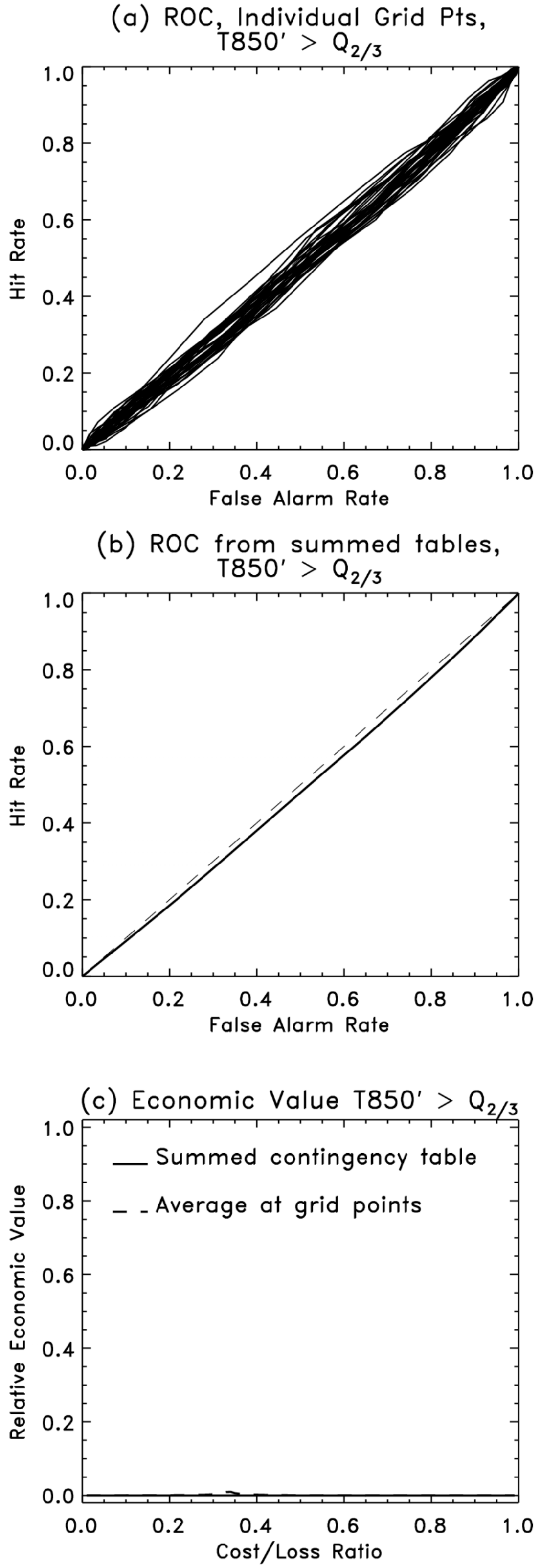

Figure 6: As in Fig. 4, but for the event of 850 hPa temperature anomaly is greater than the upper tercile of the climatological distribution.

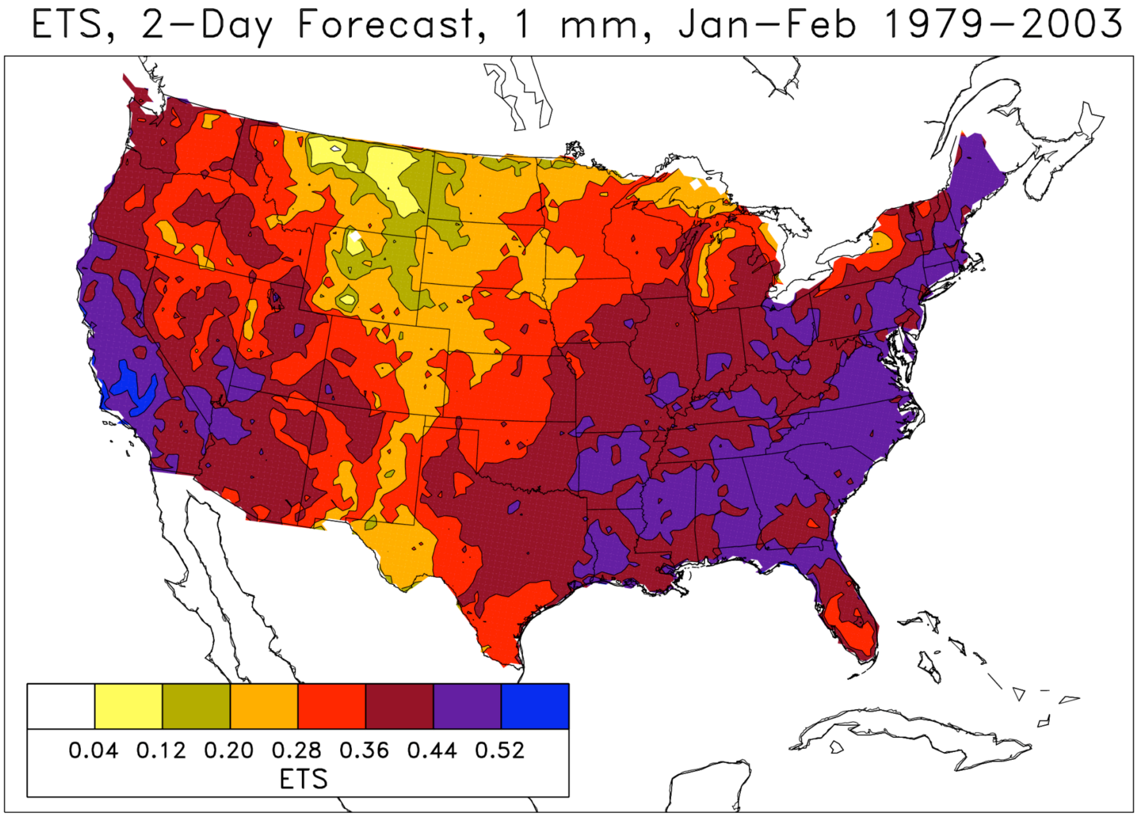

Figure 7: ETS for 1-2 day 1 mm precipitation forecasts as a function of location, using Jan-Feb 1979-2003 forecast and observational data.